A Model for Generating On-Off Speech Patterns in Two-Way Conversation

By PAUL T. BRADY

(Manuscript received February 17, 1969)

This paper describes a model that generates on-off speech patterns representative of those in experimental two-way telephone conversations. The model assumes a conversant to occupy one of three speaking or one of three silent states. Transitions among the states arc determined by Poisson processes governed by six parameters (one for each state). The validity of the model is tested by comparing the model computer simulation of 16 conversations with 16 real conversations. Cumulative distribution functions are compared for ten events (such as talkspurts, pauses, mutual silences, and so on) defined on the speech patterns. The model yields good fits to all events except "speech before interruption;" when an interruption occurs, a model speaker tends to interrupt the other's talkspurt later than a real speaker does.

Theoretical behavior of the model is also studied. All events consist of concatenations of exponentially distributed "state durations," even though most events are not themselves exponential. For some purposes, the exponential distribution is a satisfactory empirical fit to talkspurts, but not to pauses. Possible applications of the model include studying people's motivations to talk and fall silent on different circuits, and predicting statistical behavior of voice operated devices on the circuits.

I. INTRODUCTION

1.1 *Applications of the Model*

A model for generating on-off speech patterns in two-way conversations may have two uses:

(*i*) It may provide insight on the dynamic processes which determine when a person talks or is silent. For example, the model proposed here allows a person to be in one of six states, depending on whether he is talking, listening, or both conversants are talking, and so on. Each state is associated with a parameter which could be interpreted as a "motivation" for either starting to talk or falling silent. As a subject talks over different experimental conditions, changes in the "motivation parameters" might be correlated with changes in subjective opinion of the circuit.

(*ii*) The model may predict the statistical behavior of voice-operated devices (such as echo suppressors, voice-switched amplifiers) as the circuits are changed. One alternative to using a model is to have people talk over different circuits and study the circuit behavior. This is often unsatisfactory because too much data may be required to isolate the effects of a particular circuit change. Another alternative to a model is to record an experimental prototype conversation and then play it over different circuits. This is also usually unsatisfactory because the conversants cannot react to circuit changes; their behavior remains the same. A model has the advantage of keeping the statistical structure of the "conversants" unchanged while allowing them to react as the circuit parameters are varied.

A model of on-off speaking patterns is not a new concept. The design of Time Assignment Speech Interpolation* was aided by the use of a number of one-way (that is, single speaker) models in parallel to simulate speech from many subscribers.[1] Jaffe, and others, have proposed a simple two-way Markovian model intended to study the speech behavior of psychiatric patients.[2] H. W. Gustafson of Bell Telephone Laboratories has proposed some improvements on the Markovian model to allow better prediction of speaker alternations.[3] The author has twice suggested a model; the first, with Mrs. N. W. Shrimpton (unpublished work), suggested a simple exponential fit to basic events such as talkspurts, and the second used a queueing system of "ideas" and "utterances" to yield a more complex model for talkspurts.[4]

The model proposed in this paper was developed after considering a large body of data from experimental two-way conversations conducted on telephone quality circuits containing no transmission delay or other degradations (See Table II footnote and Ref. 5).† To evalu-

* TASI is essentially a bank of voice-operated switches which may disconnect a subscriber from a channel when he is not talking to permit a talking subscriber to use the channel.

† Reference 5 describes an extensive statistical analysis of speech patterns in 16 conversations, and defines many "events," such as "talkspurt," "alternation silence," "pause in isolation," and so on. Average and median lengths for the events are tabulated, and cumulative distribution functions are included. The present paper assumes prior knowledge of Ref. 5. Notice that "event" is used to mean an interval of time, such as the interval of a talkspurt, and does not mean the occurrence of a probabilistic phenomenon such as the arrival of a pulse.

ate the model, we shall compare its simulation of the 16 conversations with data from the real conversations.

1.2 *Relation of Model to Speech Detector*

A speech detector is a rule which transforms speech into on-off patterns. Speech detectors are usually designed for specific needs, and vary considerably in their specifications. If a model is fit to one speech detector's output, then the model cannot be expected to be valid for all other detectors; but with minor changes, it may be adaptable to many of them.

The author's speech detector has previously been documented, and is described briefly here.[5,6] An initial hardware detector, with virtually no "pickup" and "hangover," yields a pattern of "spurts" and "gaps," after segmenting the speech into 5 ms intervals. All spurts $\leqq$ 15 msec are presumed to be noise and are rejected (for throwaway); then all gaps $\leqq$ 200 ms are filled in, as they were probably stop consonants or other minor breaks in continuous speech. The final on-off pattern contains, by definition, "talkspurts" and "pauses." No talkspurt can be $\leqq$ 15 ms; no pause can be $\leqq$ 200 ms. The model described here therefore generates talkspurts $\geqq$ 20 ms and pauses $\geqq$ 205 ms.

The speech patterns from a speech detector are strongly influenced by choice of threshold. In this study, the Ref. 5 data taken with a -40 dBm threshold were used as a basis for the simulated conversations.

1.3 *Goals of This Paper*

The remainder of this paper is divided into two main parts. Section II describes the model and illustrates its empirical behavior by comparing its output with real conversations. The question considered is: "Can this model generate patterns statistically similar to those of a randomly selected conversation?" We do not present data on applications such as determining differences among speakers or studying the behavior of a single speaker as he engages in various tasks. Future work is planned to investigate these problems.

Section III is a mathematical analysis of the model's behavior. From this analysis, one can gain an intuitive feeling of the model behavior, and acquire insight into the manner in which the two speakers interact. For a basic treatment of the model, however, Section III may be omitted. Section II assumes an elementary knowledge of probability theory; Section III requires some background in stochastic processes.

II. MODEL DEFINITION AND EMPIRICAL BEHAVIOR

2.1 *The Model*

2.1.1 *One-Port versus Many-Port Model*

Consider speakers A and B to be engaged in conversation. We shall model only speaker A's behavior and make no attempt to include B's behavior in the formulation. That is, speaker B's patterns are regarded only as they appear to A. It may be that B is really talking, but A does not receive him because of a blocking on the transmission line. Or, B may be delayed, and A may be receiving B's previous speech when in fact B is presently silent. We shall designate our model as a "one-port" model, since only one port, that is, A's side, is formulated. To use the model, it could be connected to anything, such as another one-port model, or a one-port model connected via a transmission delay, or several one-port models as in a conference circuit. (It may be invalid to assume that speaker A can be modeled the same way in a conference as in conversation with a single other speaker, but the model does at least allow such a connection to be formulated.)

In a many-port model, the entire system is modeled, with the drawback that a separate structure is required when special circuits are inserted between speakers. In addition, a one-port model leads to a description of each speaker, while if a many-port model is used, and a real conversation between A and B differs from one between A and C, it may not be possible to ascertain the change in speaker A's behavior. All we know is that the pair A-B is different from the pair A-C.

2.1.2 *Description of the Model*

Speaker A is either talking or silent, and he views B as either talking or silent. As Fig. 1 shows, in the simplest case four states occur at A's side. A is talking in the upper (shaded) half; B is talking in the right half. In considering transitions from state to state, as shown by the arrows, we apply the restriction that the two speakers cannot change their states at precisely the same time. Thus, in Fig. 1, diagonal crossings are prohibited.

Preliminary work with the Fig. 1 model showed that it was inadequate, especially in predicting events surrounding double talk. A natural extension of Fig. 1 is to expand each state into two states, the dichotomy decided by the previous state. Figure 2 illustrates the resulting 8-state model. Consider for example "A talks, solitary": to

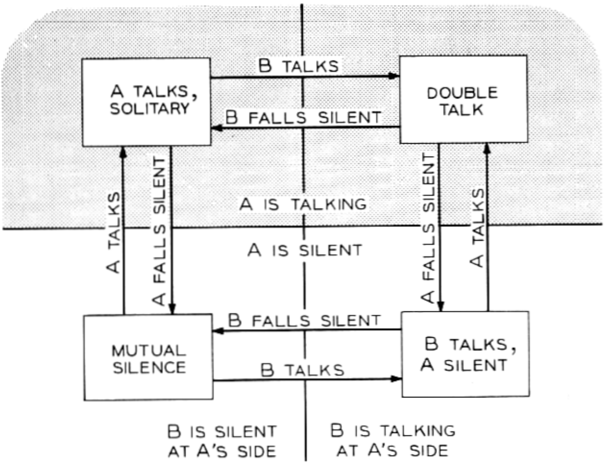

Fig. 1 — A four-state speech pattern model for speaker A. The shaded area indicates A is talking.

get to this state, either both speakers previously were silent or both were talking.

Figure 3, which is a reduction of Fig. 2, shows the model that the author has chosen to use. The upper left and lower right quadrants have been collapsed back to one state; simplicity has been gained at the expense of some loss of precision in modeling speech patterns.

Allowable state transitions are indicated on Fig. 3. There is no attempt to control B's behavior; he starts and stops talking in his own manner. Notice that his state changes cause horizontal transitions.

Vertical transitions are determined by A. If he is talking, he stops when a "fall silent pulse" occurs (Gustafson's terminology), and if silent he starts when a "start talking pulse" occurs. We call these $\beta-$ and α-pulses, respectively. These pulses are a result of Poisson processes,* so that, for example, if A is talking and B is silent (A solitary talk state), he stops talking in the next dt sec with probability $\beta_{sol}^{A} \cdot dt$.

For notation, the subscript on β, the fall silent parameter, describes the present state, while the subscript on α, the start talking parameter, denotes the event that will occur if the pulse occurs. The superscript refers to A or B. The six values for β and α are denoted β_{sol}, β_{ted}, β_{tor}, α_{pse}, α_{alt}, and α_{int} (See Fig. 3) in which the abbreviations mean solitary, interrup*ted*, interrup*tor*, pause, alternate, and interrupt.

* Poisson processes are tutorially discussed by Cox and Smith.[7]

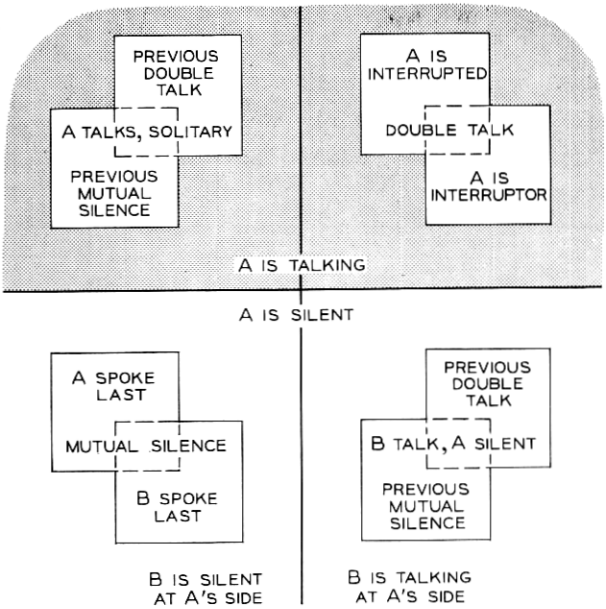

Fig. 2 — An eight-state model in which each state of Fig. 1 is divided into two states.

It is very important to understand the nature of the α or β parameters. They are not probabilities. However, if any α or β is multiplied by dt (for example, $dt = 0.005$ s), then αdt is the probability that A will leave the corresponding silence state "of his own volition" during the next dt seconds. (He may of course also be forced out of the state by B's action.) The αdt's and βdt's are "transitional" probabilities and do not represent the probability of being in each state. These "state" probabilities must be solved for, and can at times be difficult to obtain; they must consider the interaction of speaker A with his correspondent B. This is more fully discussed in Section III.

The α' and β's have a more appealing physical interpretation than just probability parameters. If some $\alpha = 2.5$, this implies that there is an "input stream" of α-pulses trying to drive A out of his state; the pulses occur at random times but at an average rate of 2.5 pulses per s, or with an average between-pulse interval of 1/2.5 s. The units of α and β are pulses per second.

In general, none of the α's or β's is time dependent, so that the duration of occupation of a state has no effect on the value of that state's

α or β. An exception is that when A becomes silent, all α's are zero for 205 ms (so that only horizontal transitions can occur), after which time they resume their model values, and when A starts to talk all β's are zero for 20 ms. This guarantees that all silences are > 200 ms, and talkspurts are > 15 ms. (If an α-pulse occurs at the 210th ms, a 205 ms interval has occurred for that state, and the remaining 5 ms are assigned to the new state.)

A summary of the assumptions made in the model is:

(*i*) At any instant of time, A exists in one of six possible states.

(*ii*) A's talk-silence behavior is governed by Poisson processes, whose parameters are functions of the state A is in, but not of the length of time in the state (except for the previously noted minimum event length requirement).

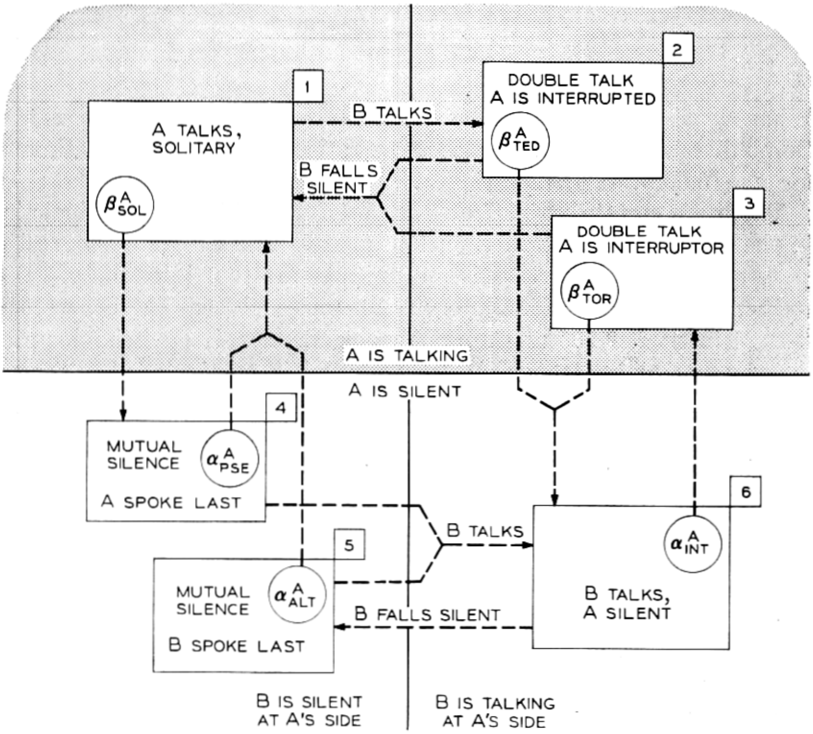

Fig. 3 — The six-state model used in this study. Vertical transitions are a result of Poisson processes at A's side. Horizontal transitions, resulting from B, are in A's external environment and are not generated by A's model.

(*iii*) The speakers cannot both change their speaking status at precisely the same instant of time.[†]

2.2 *Extracting the Model Parameters*

The six parameters for each of 32 speakers engaged in 16 conversations were derived in a very simple way: transition frequencies from state to state of the Fig. 1 model were counted by a brute-force stepping through each conversation.

To illustrate the process, recall that each person's speech is coded into 5 ms on-off intervals. Say that speaker A is in the solitary talk state (state 1, Fig. 3). β^A_{sol} can be found from a frequency count of A's falling silent from the state. Thus,

$$\beta^A_{sol} \cdot dt = \beta^A_{so:} \cdot (0.005) = \frac{\text{number of times } A \text{ falls silent from state}}{\text{number of times } A \text{ is in state, including numerator of this fraction}}. \tag{1}$$

Whenever A is in a state, his behavior can be regarded as a succession of Bernoulli trials, in which case the above ratio is a best unbiased estimator for $\beta^A_{sol} \cdot (0.005)$, and hence of β^A_{sol}.

There are certain "trials" or 5 ms intervals which are not included in the frequency count. If A just begins to talk, he cannot leave the state until the talkspurt > 20 ms, or there are four intervals of 0.005 s; therefore, the first four intervals are not included. In silences, the first 21 intervals (205 ms) are not included. Also, if B's behavior produces a horizontal transition, this interval is not included, although the intervals up to that one are counted. The rare intervals containing both a horizontal and vertical transition are counted as vertical transitions. Table I is a list of the α and β values for all 32 speakers in 16 conversations.

2.3 *Testing the Model*

2.3.1 *Method of Testing*

To investigate the behavior of the model, a Monte Carlo simulator generated a model conversation of any desired length in a form which could be analyzed by Mrs. N. W. Shrimpton's speech analysis

[†] In the author's simulation, they cannot both change status in the same 5 ms time slot. This does occur in the author's data from the speech detectors, but it is very rare.

TABLE I—α AND β VALUES FOR 32 SPEAKERS, 16 CONVERSATIONS

β_{sol}	β_{ted}	β_{tor}	Conversation		Speaker	α_{pse}	α_{alt}	α_{int}
1.261455	3.411029	2.116402	1	FEMALE	1	1.831701	1.818579	0.563380
0.996458	2.225755	1.921885	1		2	2.457002	1.285416	0.588403
0.836082	2.018229	2.496434	2		1	3.610856	1.586914	0.319719
0.614610	1.817570	1.236264	2		2	3.177857	0.735383	0.193737
0.755323	2.134756	1.428163	3		1	1.391631	0.913121	0.354644
1.334627	1.687075	1.746107	3		2	1.701480	1.061364	0.426932
0.963705	2.773246	3.626714	4		1	1.497570	0.840841	0.281668
0.920078	1.170079	3.011515	4		2	1.514693	0.575869	0.139994
0.494011	1.156069	2.501303	5	MALE	1	2.033156	1.644157	0.107033
0.229878	1.290323	2.326664	5		2	3.231441	0.812301	0.183579
0.487893	1.976664	2.629389	6		1	1.880036	1.282691	0.406757
0.606033	1.421352	0.921261	6		2	3.042993	2.743970	0.688970
1.148240	2.318487	2.191781	7		1	1.442522	0.990402	0.123400
0.945180	3.740648	2.219321	7		2	1.271811	0.805795	0.205679
0.733397	2.586904	2.464066	8		1	2.099086	0.511809	0.088622
0.710458	1.932367	3.339192	8		2	2.207059	0.588446	0.199562
0.773362	1.574803	2.679831	9		1	1.606426	0.925181	0.230204
0.660433	1.994681	2.577710	9		2	1.742712	0.873908	0.217066
0.823098	4.733728	1.898734	10		1	1.790183	0.746016	0.118188
0.670252	6.010929	4.713805	10		2	1.646938	0.462535	0.103896
0.444459	1.107595	3.638569	11		1	2.203568	2.091714	0.417177
0.372235	1.805869	2.231405	11		2	3.431840	0.867276	0.275168
0.569062	1.234167	0.665083	12		1	1.260504	0.910657	0.271548
0.755803	1.996370	1.760921	12		2	1.784675	0.525237	0.239636
0.797666	1.438849	1.886792	13	FEMALE	1	1.798942	0.988468	0.275441
0.693121	1.790580	2.906574	13		2	3.064182	1.193967	0.260818
0.768697	1.940035	1.702128	14		1	1.563188	1.446204	0.356520
0.779857	2.012072	1.672862	14		2	2.225986	0.989827	0.116141
1.075338	1.115880	2.577320	15		1	2.040816	0.521993	0.206693
0.866694	1.822600	1.962533	15		2	2.952029	0.979129	0.474661
0.652447	2.364865	3.157895	16		1	2.615694	0.292312	0.134417
1.671175	2.427184	2.173913	16		2	4.990944	1.389631	0.343234

program. (The output of the program is illustrated in Ref. 5 and some of it is shown here.) The general procedure was to extract parameters from a real conversation and then simulate a conversation of 20 minutes duration. The original conversations were between 7 and 10 minutes long, but the simulated ones were longer to better estimate the true theoretical behavior. (Economic considerations prohibited simulations significantly longer than 20 minutes.)

If we could regard the two conversations of a real-simulated conversation pair as independent samples from two populations (or the same population), then classical statistical tests (such as *t*-test on means) would be appropriate. Unfortunately, the simulated conversations were derived from measurements of the real conversations, and standard tests no longer apply. For example, say that the talkspurt average lengths were very close for real and simulated conversations. With independent samples, this would suggest a good fit, but it may be that we have forced a good fit by setting simulated parameters equal to measured parameters of real speech.

Instead of using statistical tests, we define a "fit parameter," or *FP*, to indicate the correspondence between real and simulated events. This correspondence is examined for three quantities: average lengths of the events, cumulative distribution functions (cdfs) of the events, and rate of occurrence (for example, number of talkspurts per second). These three quantities are not independent; for example, a good fit of the cumulative distribution function (cdf) implies a good fit to the average (but the converse is not true). In assessing a good or bad fit of the model to the speech data, the fit parameters are not treated as yielding three independent pieces of information, but rather as representing three viewpoints of the goodness of a fit problem.

Table II is a list of the ten events. Two events, double talks (3) and mutual silences (4), merit a brief comment. In the experimental conversations, with no circuit degradation or delay, these events are identical for both speakers. However, for consistency with the other events, comparisons of the fit parameter are made twice, once for each speaker. This causes some redundancy in the tabulated comparisons of events 3 and 4 in Tables III through V.

2.3.2 *Average Lengths*

For average lengths, we define

$$FP_{(x)} \text{ (fit parameter)} \equiv \frac{\langle x \rangle_{real} - \langle x \rangle_{sim}}{\left(\dfrac{\sigma_{real}^2}{n_{real}} + \dfrac{\sigma_{sim}^2}{n_{sim}} \right)^{\frac{1}{2}}}, \tag{2}$$

TABLE II—CATEGORIZED SPEECH EVENTS

Number	Event*
1	Talkspurt
2	Pause
3	Double talk
4	Mutual silence
5	Alternation silence†
6	Pause in isolation
7	Solitary talkspurt
8	Interruption
9	Speech after interruption
10	Speech before interruption

* For definition of "event" see Ref. 5.

† In Fig. 6 of Ref. 5, the alternation silences illustrated in the sample patterns are all incorrectly labeled. The *A*'s and *B*'s are transposed.

that is, the normalized difference between the means. If the observations were independent and from the same population, $FP_{\langle x \rangle}$ would be normal, $\mu = 0, \sigma^2 = 1$. Independence is violated here, but we still can regard FP as an indication of similarity, and arbitrarily regard the fit as "bad" if $| FP_{\langle x \rangle} | > 1.96$. Table III lists $FP_{\langle x \rangle}$ for 10 events, 32 speakers. (See Ref. 5 for tabulated average lengths of real speech events.)

2.3.3 *Cumulative Distribution Functions*

In a two-sample Kolmogorov-Smirnov test, in which n_1 observations are made for one sample and n_2 for the other, the test statistic D is the maximum vertical discrepancy (absolute value) between the cumulative distribution functions for the two samples. If both n_1 and n_2 exceed 40, the identical population hypothesis is rejected at 0.05 level (see p. 131 of Ref. 8) if

$$D > 1.36 \left(\frac{n_1 + n_2}{n_1 n_2} \right)^{\frac{1}{2}}. \tag{3}$$

Again, in our data the samples are not independent, and although n_{sim} almost always exceeds 40, n_{real} often does not exceed 40 for those events surrounding interruptions. Nevertheless, we define

$$FP_{cdf} \equiv \frac{D}{1.36} \left(\frac{n_1 n_2}{n_1 + n_2} \right)^{\frac{1}{2}}. \tag{4}$$

If $FP_{cdf} > 1$, the fit will be considered bad. Table IV lists FP_{cdf} for 10 events and 32 speakers.

Comparative plots of cdf_{real} versus cdf_{sim} for all 10 events and all 32 speakers were generated. The curves for events 1 and 10 for speaker

TABLE III—$FP_{\bar{x}}$ FOR 32 SPEAKERS IN 16 CONVERSATIONS

Conversation	Speaker	Event									
		1	2	3	4	5	6	7	8	9	10
1	1	0.81	1.22	0.42	0.18	1.05	0.49	1.66	0.12	−0.24	1.50
	2	−0.06	−0.57	0.42	0.18	−0.35	−0.41	−0.34	0.50	0.87	−2.44*
2	1	0.54	0.65	0.14	0.44	0.81	−0.86	0.66	−0.40	−0.40	−1.07
	2	0.23	−0.38	0.14	0.44	1.25	−0.54	2.12*	−1.91	0.01	0.17
3	1	1.60	−1.12	1.28	−0.23	1.88	−1.05	3.67*	0.87	0.94	−2.58*
	2	0.67	1.26	1.28	−0.23	1.37	−2.50*	0.96	0.63	−0.38	−2.83*
4	1	−0.39	0.79	0.10	1.35	2.56*	−0.60	−0.03	0.54	−1.67	−2.62*
	2	−0.04	−0.42	0.10	1.35	1.08	−0.96	1.23	0.61	0.24	−2.56*
5	1	−1.53	−0.24	−0.15	0.46	0.26	−0.73	−2.35*	−0.48	−0.40	−1.20
	2	−0.63	−1.07	−0.15	0.46	0.19	0.56	−0.32	0.06	0.70	−2.25*
6	1	0.40	0.18	−0.04	0.34	−0.54	−0.50	1.67	−1.33	−1.58	−3.04*
	2	−0.44	0.37	−0.04	0.34	0.66	0.48	2.24*	−0.06	1.23	−0.64
7	1	−0.82	0.35	−0.44	1.06	1.97*	1.04	−.53	0.57	−1.17	−1.42
	2	0.24	−0.17	−0.44	1.06	0.30	−1.96*	1.31	−1.38	0.22	−1.19
8	1	−0.43	0.04	0.53	0.72	0.96	−0.15	0.58	−1.38	0.97	−2.63*
	2	1.30	0.18	0.53	0.72	0.26	0.18	2.02*	−1.13	−0.05	−0.07
9	1	0.58	0.62	0.39	0.03	−0.60	0.14	1.32	0.66	0.02	−0.47
	2	0.89	−0.27	0.39	0.03	−1.47	1.20	1.45	−1.33	−1.85	−0.23
10	1	0.39	0.25	0.15	0.93	−0.13	0.71	0.88	−0.17	0.19	0.87
	2	0.19	−0.08	0.15	0.93	−0.70	1.05	−0.23	0.16	−1.01	−0.39
11	1	0.81	−0.15	0.13	0.82	1.80	−0.67	2.73*	−0.84	−0.28	−0.50
	2	−0.52	0.26	0.13	0.82	0.96	−1.43	0.71	−0.53	−0.68	−2.47*
12	1	1.20	0.28	0.61	0.21	−0.77	−1.39	1.80	1.25	−0.03	−0.35
	2	−0.01	0.04	0.61	0.21	2.39*	0.43	−0.58	1.06	−0.87	−1.12
13	1	0.57	−0.78	−0.05	0.63	1.52	0.12	1.52	−0.61	−1.05	−1.66
	2	0.24	0.57	−0.05	0.63	−0.17	−0.23	1.62	−0.79	−0.21	−1.29
14	1	0.58	0.05	−0.64	0.23	0.75	−0.54	0.92	0.17	0.47	−3.12*
	2	0.39	−0.07	−0.64	0.23	0.45	−0.70	1.41	−0.87	0.66	−1.84
15	1	0.19	0.14	0.69	0.09	1.10	−1.16	0.55	−0.24	−0.35	−0.35
	2	0.33	−0.20	0.69	0.09	0.98	−0.85	0.96	−0.57	0.40	−1.60
16	1	0.05	−0.42	−1.26	−0.20	2.05*	−0.81	1.21	−0.39	1.47	−2.42*
	2	0.34	0.24	−1.26	−0.20	−0.05	−1.31	−0.00	−1.20	−1.49	−1.37
Total number of "bad" fits		0	0	0	0	4	2	6	0	0	11

$$FP_{\bar{X}} \equiv \frac{\bar{x}_{real} - \bar{x}_{sim}}{\left(\dfrac{\sigma_{real}^2}{n_{real}} + \dfrac{\sigma_{sim}^2}{n_{sim}}\right)^{\frac{1}{2}}}$$

The fit is considered "bad" if $|FP| \geq 1.96$; however, this must not be regarded as a statistical test. Bad fits are marked with asterisks. For typical values of $\bar{x}_{real}$ see Table IV of Ref. 5. Conversations 5 through 12 involve men, the rest involve women.

TABLE IV—FP_{cdf} FOR 32 SPEAKERS IN 16 CONVERSATIONS

Conversation	Speaker	Event									
		1	2	3	4	5	6	7	8	9	10
1	1	0.53	0.56	0.46	0.63	0.68	0.46	0.89	0.66	0.39	0.89
	2	0.59	1.44*	0.46	0.63	0.75	0.59	0.33	0.40	0.41	1.47*
2	1	0.87	0.74	0.47	0.51	0.53	0.78	0.93	0.34	0.49	1.07*
	2	0.39	1.17*	0.47	0.51	0.75	0.58	0.98	0.55	0.34	0.45
3	1	0.85	0.84	0.72	0.62	1.01*	1.15*	1.24*	0.37	0.49	1.63*
	2	0.97	0.74	0.72	0.62	0.65	1.02*	0.76	0.53	0.87	1.17*
4	1	0.80	0.97	0.56	0.69	1.28*	1.05*	0.67	0.52	0.80	1.16*
	2	0.40	1.15*	0.56	0.69	0.76	1.24*	0.66	0.23	0.56	0.91
5	1	0.61	1.16*	0.64	0.62	0.81	0.86	0.67	0.38	0.59	1.20*
	2	1.21*	0.69	0.64	0.62	0.52	0.47	1.13*	0.64	0.34	1.03*
6	1	0.77	1.11*	0.62	0.66	0.76	0.63	0.70	0.59	0.94	1.23*
	2	0.78	1.12*	0.62	0.66	0.68	0.62	1.06*	0.37	0.48	1.23*
7	1	0.76	0.48	0.38	0.74	0.84	0.85	0.50	0.53	0.46	0.69
	2	0.61	0.66	0.38	0.74	1.26*	0.90	0.76	0.72	0.47	0.89
8	1	0.90	0.63	0.47	0.65	0.99	0.70	0.70	0.54	0.41	1.14*
	2	1.38*	0.89	0.47	0.65	0.97	0.88	1.69*	0.37	0.42	1.15*
9	1	0.42	0.77	0.85	0.48	0.53	0.30	0.86	0.38	0.63	0.85
	2	0.52	0.84	0.85	0.48	0.76	0.48	0.78	0.58	0.82	0.86
10	1	0.90	0.89	0.45	0.54	0.69	0.79	0.98	0.62	0.40	0.58
	2	0.72	0.57	0.45	0.54	0.66	0.44	0.61	0.63	0.57	0.52
11	1	0.81	0.42	0.84	1.08*	0.92	0.49	1.25*	0.56	0.33	0.85
	2	0.88	0.75	0.84	1.08*	1.33*	0.60	0.46	0.91	0.88	0.79
12	1	0.78	0.63	0.52	0.55	0.34	0.72	1.25*	0.62	0.54	1.07*
	2	0.27	0.64	0.52	0.55	0.97	0.42	0.34	0.57	0.64	1.06*
13											
	1	0.60	0.78	0.56	0.39	0.54	0.62	0.78	0.49	0.39	1.10*
	2	0.52	0.41	0.56	0.39	0.46	0.45	0.56	0.51	0.48	0.50
14	1	1.24*	1.15*	0.59	1.08*	0.92	0.42	1.37*	0.28	0.50	0.95
	2	0.57	0.71	0.59	1.08*	1.22*	1.11*	0.64	0.32	0.77	0.77
15	1	0.50	0.96	0.83	0.57	0.59	1.03*	0.42	0.42	0.45	0.83
	2	0.53	0.58	0.83	0.57	0.96	0.60	0.66	0.70	0.40	1.08*
16	1	0.62	0.86	0.54	0.48	0.85	0.55	0.93	0.41	0.74	0.77
	2	0.65	0.68	0.54	0.48	0.33	0.71	0.72	0.56	0.63	0.78
Total number of "bad" fits		3	7	0	4	5	6	7	0	0	15

$$FP_{cdf} \equiv \frac{D}{1.36}\left(\frac{n_{real}n_{sim}}{n_{real}+n_{sim}}\right)^{\frac{1}{2}},$$

where D is the maximum absolute vertical distance between the two cumulative distribution functions. A "bad" fit occurs if $FP_{cdf} \geq 1.0$, as marked by asterisks. Cumulative distribution functions for events for all speakers collectively (for example, all talkspurts lumped together) are shown in Ref. 5.

TABLE V—VALUES FOR $FP_n \equiv \dfrac{\text{RATE OF OCCURRENCE OF SIMULATED EVENT}}{\text{RATE OF OCCURRENCE OF EVENT IN REAL CONVERSATION.}}$

Conversation	Speaker	Events							
		1	2	3	4	5	6	7	8
1	1	1.072	1.074	1.039	1.029	1.016	1.183	1.091	1.047
1	2	0.976	0.978	1.039	1.029	0.966	0.955	0.832	1.036
2	1	1.065	1.065	1.012	1.041	1.066	1.104	1.082	0.974
2	2	0.983	0.986	1.012	1.041	1.016	0.898	0.920	1.061
3	1	0.990	0.990	0.972	1.080	0.980	0.960	0.960	1.065
3	2	1.101	1.099	0.972	1.080	0.992	1.305	1.197	0.910
4	1	1.040	1.042	0.924	1.022	1.011	1.128	1.068	0.891
4	2	0.959	0.957	0.924	1.022	0.968	0.950	0.966	1.008
5	1	0.912	0.908	0.983	0.862	0.890	0.864	0.771	1.054
5	2	0.879	0.884	0.983	0.862	0.859	0.877	0.681	0.941
6	1	1.028	1.026	1.046	0.986	1.023	1.196	1.148	1.045
6	2	0.996	0.996	1.046	0.986	0.926	1.001	0.911	1.061
7	1	1.006	1.006	0.931	1.012	1.029	1.031	1.011	0.833
7	2	0.999	0.996	0.931	1.012	0.960	1.039	1.034	0.994
8	1	0.982	0.979	0.957	1.052	1.008	1.040	0.976	0.702
8	2	1.079	1.079	0.957	1.052	1.042	1.088	1.106	1.117
9	1	1.082	1.082	0.943	1.087	1.021	1.266	1.188	0.849
9	2	1.018	1.020	0.943	1.087	1.022	1.031	1.005	1.056
10	1	1.039	1.037	0.919	1.053	1.105	1.046	1.036	0.892
10	2	1.023	1.026	0.919	1.053	1.128	0.998	1.074	0.945
11	1	1.037	1.037	1.021	1.024	1.039	0.968	0.812	1.119
11	2	1.005	1.005	1.021	1.024	1.027	1.182	1.025	0.905
12	1	1.075	1.075	1.020	1.066	1.041	1.160	1.119	1.021
12	2	1.022	1.022	1.020	1.066	1.001	1.044	0.994	1.042
13	1	1.016	1.014	1.019	1.050	1.013	1.047	1.003	0.999
13	2	1.058	1.060	1.019	1.050	1.019	1.112	1.036	1.046
14	1	1.030	1.028	0.839	1.080	1.131	1.084	1.086	0.856
14	2	1.018	1.018	0.839	1.080	1.080	1.036	1.082	0.801
15	1	1.027	1.027	0.976	1.028	0.999	1.131	1.140	0.972
15	2	1.009	1.009	0.976	1.028	1.037	1.008	0.984	1.009
16	1	0.921	0.917	0.822	1.056	1.005	0.949	1.034	0.846
16	2	1.053	1.054	0.822	1.056	1.074	1.079	1.074	0.802

Events 8, 9, and 10 occurred an equal number of times and have equal FP_n.

2 of conversation 12 were arbitrarily selected for inclusion in this paper as Figs. 4 and 5. They illustrate a good and bad fit of the cumulative distribution functions, respectively. The plotted points are not data points; they represent category intervals of 15, 20, 30, $\cdots$ 200 ms, and 1, 2, 3 s, and so on. Thus, the number of asterisks in the cumulative distribution functions of real speech, or of breakpoints in the connected curves of cumulative distribution functions of simulated speech, do not equal n_{real} and n_{sim}, respectively.

2.3.4 *Rate of Occurrence*

To compare rate of occurrence of events

$$FP_n \equiv \frac{n_{sim}/\text{length of sim conversation}}{n_{real}/\text{length of real conversation}}. \qquad (5)$$

For a good fit, FP_n should be close to 1.0. When either n is small, FP_n may be changed considerably by the addition or subtraction of even one event, and unfortunately, FP_n does not consider the absolute values of the n's in the comparison, as do the other two FP's. In addition, we have not found a statistical test which is suitable for comparing rates of occurrences of events such as our speech events. Table V therefore is included only as a listing of the values of FP_n for eight events, 32 speakers without specifying good or bad fits. (Events 8, 9, and 10 occur an equal number of times; thus FP_n is equal for the three events.)

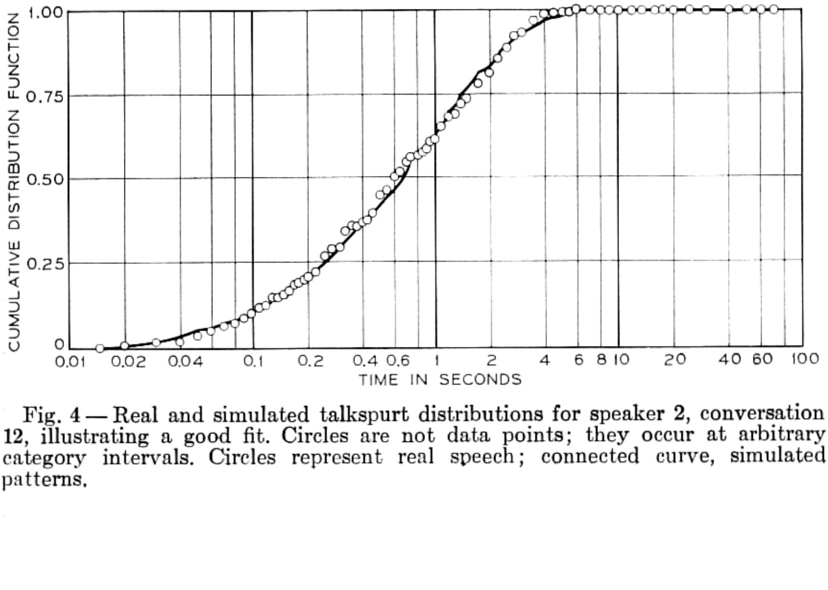

Fig. 4 — Real and simulated talkspurt distributions for speaker 2, conversation 12, illustrating a good fit. Circles are not data points; they occur at arbitrary category intervals. Circles represent real speech; connected curve, simulated patterns.

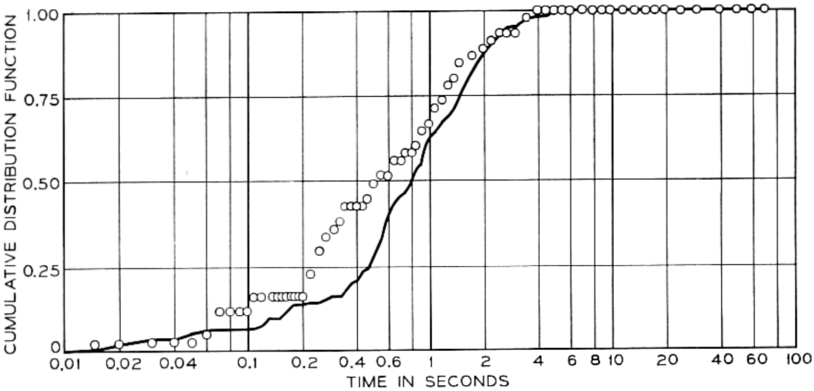

Fig. 5 — Real and simulated speech before interruption distributions, speaker 2, conversation 12, illustrating a bad fit.

2.4 *Discussion*

2.4.1 *Goodness of Fit*

In this study, the model is regarded as successful if it can match the distributions and rates of occurrence of the ten events listed in Table II. To see how well this criterion is met, the three *FP*'s are considered separately.

First, notice in Table III that several columns (events) have no "bad" fits for average lengths. Now, in 32 trials of a legitimate 0.05 level test, we would expect about 1.5 failures; the six columns with no failures represent 160 trials (192 trials minus the 32 redundant trials for events 3 and 4) with eight expected failures. The lack of failures tends to rule out the $N(0, 1)$ distribution of entries in a column. Further, inspection of column 2, for example, reveals a variable which is apparently not $N(0, 1)$; 28 out of 32 (87.5 percent) of observations are within ± 1, as opposed to 68 percent for $N(0, 1)$. This substantiates our earlier statement that use of the *FP* does not constitute a legitimate statistical test.

Without regard to statistics, however, the Table III data do show a "good" correspondence between real and simulated (that is, model predicted) averages for all events except 10 (speech before interruption) and possibly 7 (solitary talkspurts). Looking now at Table IV, the model is clearly inadequate for event 10, but event 7 is not much worse than the rest. The Kolmogorov-Smirnov test (using $FP_{cdf} \geqq 1$ as failure criterion) is powerful and would be a severe test if statistical tests were valid; for this reason, we regard the cumulative distribution functions fits as generally successful (except for event 10), but with room for improvement. Further, some fits are remarkably close, such as the talkspurt cumulative distribution functions for conversation 12 speaker 2 (see Fig. 4).

The rate of occurrence ratios in Table V are generally close to 1.0; there are a few scattered discrepancies but the model does not appear to have serious problems in generating a realistic number of events.

Regarding individual conversations, a study of Tables III and IV shows no tendency for the model to fail on particular speakers. The speakers in conversation 3 have a few more failures than seems normal, but even here the model exhibits "acceptable" fits for most events. The model appears equally valid for men (conversations 5 through 12) as for women (1 through 4 and 13 through 16).

2.4.2 *Conversational Behavior*

The failure of the model in predicting speech-before-interruption intervals may shed some light on the behavior of the subjects. Table III shows that the simulated intervals are too long; real people tend to interrupt sooner than predicted by simulation. This may be a question of reaction time. The model assumes that the instant A (who is silent) hears B begin to speak, A is immediately in a "listen to B" state and will speak only if he wishes to interrupt. In reality, A may require some time—perhaps 200 ms—before he adjusts to the presence of B's speech; in the meantime A's speech may not be intended as an interruption. A more sophisticated model might in fact assume the existence of a short delay in A's reception of B.

The numerical values of the α's and β's (Table I) also provide clues to behavior. The absolute values are a little hard to interpret since they are so closely related to the design of the author's speech detector. But notice that α_{alt} is less than α_{pse} for each speaker, confirming our intuitive belief that a person is more likely to resume talking after a pause he generates than after a pause the other party generates. This also justifies having two different states for A in which B is silent; the model would certainly deteriorate if the states were merged to one with an "averaged" α parameter.

Considering β_{tor} and β_{ted}, there is no consistent difference; $\beta_{ted} > \beta_{tor}$ for 15 of the 32 speakers. Thus, 15 (about half) of the subjects are more likely to terminate double talking if interrupted than if they are interruptors. A simpler model might merge these states, but serious errors might result for some subjects, since β_{ted} is often considerably different from β_{tor}.

The lowest of the α's is predictably α_{int}. A person is less likely to start talking when his correspondent is talking than when he is silent.

The precision of the α's and β's merits some attention. Because these quantities were measured over a person's entire conversation, they are not statistical estimators, but exact measures, correct to six figures. If you wish to regard the conversation as a sample of a larger population, however, you could regard an estimated value of α or β as a measure of a population α or β and establish confidence limits. The α's and β's were measured from Bernoulli trials where n varied from about 1000 to 40,000, depending on the conversation and parameter to be measured. Although the n's were large, the p values ($\alpha\, dt$) were generally very small, typically about 0.005; standard deviations of α or β estimates could equal about 0.1, with resulting 95 percent confidence limits of about ±0.2.

2.4.3 *Scope of the Model*

Telephone conversations usually begin with a brief but rapid interchange of short words ("hello," and so on). In many calls the calling party then assumes dominance, and then possibly the other party may dominate. Our model attempts to duplicate speech patterns using six time-invariant parameters for each speaker and cannot, except by chance, generate the alternation of dominance which often occurs in real conversations.

The model is, however, a very simple one. With only six states we are attempting to simulate the utterance patterns of a person, who is certainly not a six-state device. Simplicity is also achieved by the Markovian technique of having a person leave a state with a time-invariant probability, independent of the duration of state occupation. (The minimum pause and talkspurt lengths constitute minor violations of this philosophy, but add little to the complexity of the model.)

The real issue here is not whether such a simple model can duplicate *all* aspects of conversation behavior, but rather whether such a model is useful on its own. The author plans to test it by using it to investigate speech behavior on circuits with transmission delay; another group at Bell Laboratories is studying its applicability to circuits with switched-gain amplifiers. The ease with which the model can be simulated, plus its success in matching overall patterns, gives it the potential of becoming an important tool in the study of conversational dynamics.

It may eventually prove worthwhile to extend the model and try to get a closer match to the dynamics of conversation. One way to do this would be to increase the number of states. This might improve the fit to the "total pattern" distribution, but might require a huge number of states before a realistic "dominance alternation" occurs. Another way would be to introduce time-varying α and β parameters in the present six-state model. It appears that the development of either of these extended models (or a combination of them) would require an intensive amount of additional research.

III. MATHEMATICAL ANALYSIS

The principal goal of this section is to find theoretical distribution functions of the ten speech events in Table II. A complete analysis of the Fig. 3 model is not possible, but it is possible to analyze a simplified model and extend the results. For analysis, the model must be connected to another speaker. Section 3.1 considers speakers A and B to be directly connected with no minimum pause and talkspurt restrictions. Section 3.2 introduces these restrictions, to make the model match the author's speech detector. Section 3.3 considers an exponential approximation to talkspurts and pauses, and Section 3.4 discusses the effects on the analysis of introducing special circuits between subjects (transmission delay, echo suppressors).

3.1 *Direct Connection of Two Speakers*

Let the speech pattern model (Fig. 3) for speaker A be directly connected with one for speaker B. The entire A-B system thus exists in six states, since each state for A can be shown to correspond to a unique state for B. If all α's are forced to zero in the first 200 ms of silence, then a 200 ms minimum pause restriction is achieved; if β's are zero for 15 ms of talking, a 15 ms minimum talkspurt is achieved. In this section, we do not use these restrictions; we regard all α's and β's as time invariant. Because each state is terminated by a Poisson pulse from either A or B, the entire system is Markovian and the duration of each state has an exponential distribution.

This is illustrated, for example, by state 5. A will leave state 5 of his own volition in dt seconds with probability $\alpha_{alt}^{A} \cdot dt$. State 5 for A corresponds to state 4 for B; hence, B causes A to leave state 5 with probability $\alpha_{pse}^{B} \cdot dt$. A remains in state 5 with probability $1 - \alpha_{alt}^{A} \cdot dt - \alpha_{pse}^{B} \cdot dt$.* State 5 is thus terminated by a Poisson process with parameter $(\alpha_{alt}^{A} + \alpha_{pse}^{B})$; its duration is exponentially distributed with that parameter (see p. 154 of Ref. 9).

The appendix shows that even if only those events are considered in which A happens to terminate a state, these events are also exponential

* Cox and Smith give an expository treatment of this kind of analysis.[7]

with the parameter equal to the sum of the A and B "exit" parameters. For example, a "solitary talkspurt," in which A generates a talkspurt entirely within B's silence, is terminated when A leaves state 1 because of a β^A_{sol} - pulse. Nevertheless, A's solitary talkspurt is exponential with parameter $(\beta^A_{sol} + \alpha^B_{int})$ and therefore has an average length of $1/(\beta^A_{sol} + \alpha^B_{int})$. (State 1 at A's side corresponds to state 6 at B's side.)

This prediction for solitary talkspurt average lengths is well supported by simulation and is in fair agreement with actual speech data. Table VI compares the predicted average talkspurt lengths for 32 speakers with the measured averages from simulation. Only 2 out of 32 fail a 5 percent level test, which indicates that the simulator (that is, model) behaves as predicted.

Table VI also shows data from real speech. It is more appropriate to compare the real speech averages with simulated averages than with theoretical predictions, since the simulator contained the 15 ms and 200 ms minimum talkspurt and pause restrictions. Table III showed that 6 of the 32 average lengths of simulated solitary talkspurts were judged to be "bad" fits to empirical averages. In addition, a product-moment correlation of 0.91 is found for the two columns of average lengths in Table VI. A reasonably good fit is thus suggested; but Table VI shows that the real speech average exceeds the simulated average in 25 of the 32 cases. There is therefore a definite but mild tendency for the model (that is, simulator) to predict solitary talkspurts which are too short. This in no way refutes the result of the appendix, which is related only to the theoretical model.

In summary, the six-state Markovian system in this section may be solved by standard techniques. The following conclusions seem most relevant to speech analysis.

(*i*) A solution of the steady state probabilities of being in each of six states (that is, percent time in each state) may be obtained by routine solution of Markovian transition equations. This solution is not presented here because it is cumbersome, and it is not required for finding the distributions of durations of many of the states.

(*ii*) The distribution of the duration of A's being in any one of the six states is exponential with its parameter equal to the sum of the A and B parameters for leaving the state.

(*iii*) The distribution of three speech events may be immediately deduced. The events are:

(*a*) Alternation silence from B to A, in which B stops talking, there is a mutual silence, and A starts. This is distributed as the duration of state 5: exponential $(\alpha^A_{alt} + \alpha^B_{pse})$.

Table VI—Predicted and Empirical Average Solitary Talkspurt Lengths

			Simulated		Real Speech	
Conversation	Speaker	Predicted (s)	n	Ave (s)	n	Ave (s)
1	1	0.541	342	0.556	158	0.655
	2	0.641	246	0.628	149	0.606
2	1	0.971	377	0.983	183	1.035
	2	1.070	149	0.995	85	1.343
3	1	0.846	213	0.773	116	1.172
	2	0.592	380	0.566	166	0.621
4	1	0.906	374	0.969	208	0.966
	2	0.832	231	0.834	142	0.952
5	1	1.476	122	2.225*	87	1.631
	2	2.968	73	3.282	59	3.045
6	1	0.850	79	0.729	38	1.135
	2	0.987	175	0.957	106	1.259
7	1	0.739	363	0.832*	181	0.796
	2	0.936	320	0.942	156	1.064
8	1	1.072	222	1.114	117	1.192
	2	1.251	333	1.160	155	1.380
9	1	1.010	280	1.013	82	1.203
	2	1.123	231	1.025	80	1.242
10	1	1.079	356	1.071	116	1.157
	2	1.268	226	1.271	71	1.234
11	1	1.390	144	1.290	51	1.879
	2	1.267	82	1.214	23	1.467
12	1	1.237	226	1.130	77	1.405
	2	0.973	172	1.027	66	0.948
13	1	0.945	214	0.880	76	1.087
	2	1.032	282	1.012	97	1.238
14	1	1.130	336	1.121	118	1.221
	2	0.880	244	0.860	86	1.047
15	1	0.645	203	0.629	56	0.684
	2	0.932	388	0.893	124	0.992
16	1	1.004	116	0.958	18	1.335
	2	0.554	864	0.549	129	0.549

Predicted averages for A speakers (speakers 1) = $1/(\beta_{sol}{}^{A} + \alpha_{int}{}^{B})$, for B speakers = $1/(\beta_{sol}{}^{A} + \alpha_{int}{}^{A})$. Values for α's and β's were obtained from Table I. This prediction is slightly in error because of the 200 ms minimum pause requirement, as explained in Section 3.2. Significance (marked by asterisks) is at 0.05 level; $\bar{x}$ is assumed normal with mean = predicted average, σ = mean$/(n)^{\frac{1}{2}}$, since for a single observation from exponential distribution, $\sigma = \mu$. (Simulated and real speech n's are considerably different because lengths of conversations are different.) Product-moment correlation of simulated and real averages = 0.91.

(*b*) Pause in isolation, which has the distribution of state 4: exponential $(\alpha_{pse}^{A} + \alpha_{alt}^{B})$.

(*c*) Solitary talkspurt which is exponential with parameter $(\beta_{sol}^{A} + \alpha_{int}^{B})$. (State 1 also has this distribution; but A's being in state 1 does not imply a solitary talkspurt, since state 1 can be entered from double talking.)

(*iv*) Two distributions are a little more difficult, but straightforward.

(*a*) Double talk, in which states 2 and 3 are each exponential, but with different parameters. The double talk density function is an average of the two exponential density functions, each weighted by the steady state probabilities of the states 2 and 3, respectively. The resulting distribution probably resembles an exponential, but is in general not strictly exponential unless states 2 and 3 are identically distributed.*

(*b*) Mutual Silence, which is the same as in the case of double talk, but with states 4 and 5, which are each exponential.

(*v*) The distributions of the remaining events of Table II are very difficult to derive. For example, we notice that a talkspurt can consist of an infinite possible sequence of states 1, 2, and 3.† Although there are techniques for handling problems of this type, they are complicated and in this case may yield formidable analytic expressions.

Notice that for this completely Markovian system of the ten speech events of Table II, only three—alternation silence, pause in isolation, and solitary talkspurt—are strictly exponentially distributed. But all events consist of concatenations of the six states, which in turn are exponential. We could think of these states as exponential "building blocks" with which the speech events are constructed.

3.2 *Effect of Minimum Pause and Talkspurt Length*

The introduction of time-varying parameters to obtain minimum lengths for pauses and talkspurts ruins the Markovian structure of the model, and standard techniques are not applicable for solution. However, certain results are still obtainable.

First of all, in the speech model, the 15 ms minimum talkspurt requirement is included because the author's speech detector, used to collect the speech data to test the model, uses a 15 ms throwaway for noise rejection in the raw speech data; hence all measured talkspurts exceed 15 ms. Even without the 15 ms restriction in the model, most simulated speech events are much longer than 15 ms, and we can anticipate only minor errors by ignoring the minimum length in the analysis.

The 0.2 s minimum pause is harder to deal with; it is long enough to affect the results. In general, constant lengths of 0.2 s are added to exponentially distributed silent state durations. We refer to the resulting distribution as constant-plus-exponential.

In some cases, one of the state exit parameters (say β) may be zero for 0.2 s, while the other (α) remains at its usual value. Then, the first

* Averaging two density functions is not equivalent to averaging two independent exponential random variables. The latter operation yields a gamma distribution.

† Certain sequences are not possible, such as 1, 3, 2; but there are still infinitely many allowable ways A can wander among the three states before falling silent.

0.2 s is exponential with parameter α, and the probability that the interval will extend beyond 0.2 s is $e^{-0.2\alpha}$. Intervals beyond 0.2 s are constant-plus-exponential distributed, with constant = 0.2 s and exponential parameter = $(\alpha + \beta)$. Since the relative fraction of less-than versus greater-than 0.2 s intervals is known, the total distribution can be found by combining the pre- and post-0.2-s exponential segments.

These results can be used to draw the following conclusions regarding event distributions.

(*i*) Alternation silence (state 5): assuming A has not talked for 200 ms prior to the state entry;* this is exponential (α^A_{alt}) for 0.2 s, and then exponential $(\alpha^A_{alt} + \alpha^B_{pse})$.

(*ii*) Pause in isolation (state 4): these must be at least 0.2 s long, since A cannot terminate state 4 until that time. Hence, these are constant-plus-exponential distributed; constant = 0.2 s, exponential parameter = $(\alpha^A_{pse} + \alpha^B_{alt})$.

(*iii*) Solitary talkspurt tends to be exponential $(\beta^A_{sol} + \alpha^B_{int})$; most state 1 durations are unaffected by the minimum pause requirement.†

(*iv*) Double talk: states 2 and 3 distributions are completely unaffected by the minimum pause requirement, but their relative steady state probabilities may be changed somewhat, thus affecting the blend of the two density functions. The effect is probably slight, however, and the general shape of the distribution still looks very much as it did without the minimum pause length. This has the appearance of an exponential distribution, although not precisely exponential.

(*v*) Mutual Silence: this distribution was predictable without the minimum pause requirement, but it now appears to be very complex and strongly affected by the 200 ms constant. All mutual silences which are "pauses in isolation" are at least 200 ms long, and those which are "alternation silences" usually start exponentially with parameter α^A_{alt}, and after 200 ms they become exponential with parameter $(\alpha^A_{alt} + \alpha^B_{pse})$. Figure 5 of Ref. 5 clearly shows the importance of the 200 ms constant in mutual silences.

(*vi*) Remaining events are too complex to predict. Certainly, however, all talkspurts start with a 15 ms constant duration, and pauses start with a 200 ms constant duration.

3.3 *Exponential Approximation to Talkspurts and Pauses*

Exponential and constant-plus-exponential events are easy to simu-

* Ref. 5 data suggest that less than 10 percent of state 5 intervals begin within 200 ms of A's speech.

† Based on data from Ref. 5, we estimate that only about 6 or 7 percent of state 1 intervals begin with 200 ms of B's speech.

late. If one wanted to generate artificial talkspurts and pauses and was unconcerned with speaker interaction, could he use such a simplified model? We tried such a fit to the empirical talkspurt and pause distributions of the conversations described in Ref. 5.

Talkspurts were fit by a straight exponential distribution, without a 15 ms constant, in which the exponential parameter was deduced from the average event length. That is, for a particular speaker let

$$\beta_{ts} \equiv 1/\text{average talkspurt length}; \tag{6}$$

then

$$\text{Prob } (T \leqq t) \text{ (cumulative function)} = 1 - \exp(-\beta_{ts}t). \tag{7}$$

For pauses, we used a constant-plus-exponential. Let

$$\alpha_{pse} \equiv 1/(\text{average pause length } -0.2), \tag{8}$$

that is, the reciprocal average of the above 200 ms part of all pauses. Then

$$\Pr(T \leqq t) = \begin{cases} 0 \quad \text{for} \quad 0 \leqq t \leqq 0.2 \\ 1 - \exp[-\alpha_{pse}(t - 0.2)] \quad \text{for} \quad t > 0.2. \end{cases} \tag{9}$$

For comparing distribution functions, a Kolmogorov-Smirnov test was used to see if the empirical distribution function came from the particular exponential function based on β_{ts} or α_{pse}.[10] Once again, the statistical test is not strictly appropriate, since the mean of the exponential function is forced equal to the sample average. It still appears, however, to be a reasonable heuristic method to determine if the "shape" of the curve is exponential. Only four out of 32 sets of talkspurts fail the test, suggesting that the exponential model is a good approximation for talkspurts. This is in agreement with the findings of Jaffe and others.[2] None of the pauses fit constant-plus-exponential. This probably results from trying to fit one distribution to two distinctly different kinds of pause: pause in isolation, which occurs between words and is short, and the long silence which occurs when listening to the other speaker.

The good exponential fit to talkspurts might cause one to feel that the talkspurts could be modeled by a single parameter Poisson process. This would be achieved by having a single "talk" state, instead of three; once the state is entered, the speaker would ignore the other's speech and stop talking when his single parameter β-pulse occurred. Although a reasonable talkspurt fit would be achieved, other speech events, such as double talk and interruptions, would be poorly matched for most speakers. This is true because the measured values of the three β's of Fig. 3 are generally quite different, with the two double talking β's often different from each other and typically at least twice β_{sol} (see Table I). A single parameter Poisson process would incorrectly assume these β's to be equal to each other.

Why, then, do we get a good exponential fit to the general talkspurt distribution? Table II of Ref. 5, for -40 dBm threshold, shows that state 1 accounts for about 88 percent of A's talking time, so that the different β's during double talking exert only a minor effect upon the predominant state 1 single parameter Poisson process.* That is, the long and frequent state 1 intervals tend to obliterate the fine structure of the double talks.

3.4 *Connection of Two Models Over Special Circuits*

When A and B are directly connected, equations at A's side are easily written because knowledge of A's state at a random instant implies knowledge of B's state. (Once again, for simplicity, assume a Markovian model with no minimum event length requirements.) Analysis becomes very difficult when the circuit prohibits such knowledge. Two such circuits are considered here: Circuits with transmission delay and with echo suppressors.

3.4.1 *Delay*

The feasibility of transmitting two-way telephone calls over satellite circuits has generated widespread interest in the effects of transmission delay on the behavior of the conversants. We have previously dealt with a system which connected two three-state Markovian devices over a channel with transmission delay.[4,11] The following conclusions are of interest here.

(*i*) If the delay is "short" (in the order of average pause lengths or less, as occurs in cases of practical interest, where $D \leqq 1200$ ms), an exact analysis has not yet been found, and approximations are required to solve even the simple three-state system.

(*ii*) For very long delays, asymptotic system behavior of the model is obtainable; but the model is of doubtful validity since an entirely different kind of speech behavior might result from excessively long delays.

* One or the other speaker, but not both, talks for 100-24.99 (mutual silence) — 4.62 (double talk) $=$ 70.39 percent of the time; this accounts for states 1 and 6 at A's side. State 1 is occupied about half this time, or 35.20 percent. This is 88 percent of $35.20 + 4.62$, which is A's talking time.

3.4.2 *Echo Suppressors*

For our purposes, echo suppressors are devices which occasionally block the A to B or B to A (or both) transmission paths, at times depending on the interaction of the A and B speech patterns. (For further details on echo suppressors see Ref. 12.) There may also be delay, but even without delay the time dependency and uncertainty in the system is apparent and virtually prohibits formal analysis.

For both delay and echo suppressors, simulation is not difficult (the author's simulator already incorporates delay) and provides at present the only means of assessing the performance of the model.

3.5 *Summary*

The six-state model described by the author contains time dependencies which prevent formal Markovian analysis, but there is a tendency for the speech events to be formed from exponential, and in some cases, constant-plus-exponential "building blocks." Practically all of the exponential blocks or exponential parts of the constant-plus-exponential blocks have distributions with parameters equal to the sum of the A and B "exit probability" parameters; and even those events which seem exclusively a result of one speaker (such as solitary talkspurts) are in fact influenced by both speakers in a predictable way.

Although several theoretical results are obtainable, one is forced to turn to simulation for complete quantitative results. The ease by which the model is simulated helps compensate for the numerous computer runs required for studying model behavior as a function of parameter or circuit changes.

IV. ACKNOWLEDGMENT

I am grateful to Mrs. Joan Olson for writing and revising several of the computer programs required for this study.

APPENDIX

Distribution of a State Terminated By a Particular Speaker

This appendix is a derivation of the result stated in Section 3.1, that if, for example, one considers only those state 1 intervals terminated by A, these will have an exponential distribution with parameter $(\beta^A_{sol} + \alpha^B_{int})$. For shorthand, we call the parameters β and α. Let state 1 begin at time $t = 0$. The joint probability that it is t s long and terminated by A is:

$$\text{Pr (terminated by } A \text{ in } t,\ t + dt) = e^{-(\alpha+\beta)t} \cdot \beta\, dt; \qquad (10)$$

that is, neither an $\alpha-$ or $\beta-$ pulse can occur in $(0, t)$, and one β-pulse must occur in $(t, t + dt)$. Integrating equation (10) over all t,

$$\text{Pr (state terminated by } A \text{ at any time)}$$

$$= \int_0^\infty (10)\, dt = \beta/(\alpha + \beta), \qquad (11)$$

as it should. We desire the conditional probability that state 1 ends in $(t, t + dt)$ given that it is terminated by A. By Bayes' rule,

$$\text{Pr (state ends in } (t,\ t + dt) \mid \text{terminated by } A)$$

$$= \frac{\text{joint Pr (state ends in } (t,\ t + dt) \text{ and is terminated by } A)}{\text{Pr (terminated by } A)}$$

$$= \text{equation (10)/equation (11)} = e^{-(\alpha+\beta)t}(\alpha + \beta)\, dt, \qquad (12)$$

which is recognized as an exponential density function with parameter $(\alpha + \beta)$.

REFERENCES

1. Fraser, J. M., Bullock, D. B., and Long, N. G., "Overall Characteristics of a TASI System," B.S.T.J., *41*, No. 4 (July 1962), pp. 1439–1454.
2. Jaffe, J., Cassotta, L., and Feldstein, S., "Markovian Model of Time Patterns of Speech," Science, *144*, No. 3620 (May 15, 1964), pp. 884–886.
3. Gustafson, H. W., "Model for the Analysis of Talkspurt and Silence Durations in Conversational Interaction," Proc. 77th Annual Conv. Amer. Psychological Assn., *44*, Part I (1969), pp. 43–44.
4. Brady, P. T., "Queuing and Interference Among Messages in a Communication System with Transmission Delay," Ph.D. Thesis, Department of Electrical Engineering, New York University, June 1966.
5. Brady, P. T., "A Statistical Analysis of On-Off Patterns in 16 Conversations," B.S.T.J., *47*, No. 1 (January 1968), pp. 73–91.
6. Brady, P. T., "A Technique for Investigating On-Off Patterns of Speech," B.S.T.J., *44*, No. 1 (January 1965), pp. 1–22.
7. Cox, D. R., and Smith, W. L., *Queues,* Methuen: London, 1961.
8. Siegal, S., *Nonparametric Statistics for the Behavioral Sciences,* New York: McGraw-Hill, 1956.
9. Cox, D. R., and Miller, H. D., *The Theory of Stochastic Processes,* New York: Wiley, 1965.

10. Hoel, P. G., *Introduction to Mathematical Statistics,* New York: Wiley, 1962.
11. Brady, P. T., "A Stochastic Model of Message Interchange on a Channel With Transmission Delay," IEEE Trans Commun. Techniques, *15,* No. 3 (June 1967), pp. 405–412.
12. Unrue, J. E., "Echo Suppressor Design Considerations," IEEE Trans. Commun. Techniques, *16,* No. 4 (August 1968), pp. 616–624.